# SECONDARY USE OF RESEARCH DATA IN THE EU: COMPLEX INSTITUTIONAL APPROACH

Jakub Harašta / Matěj Myška

Research fellow, Masaryk University, Faculty of Law, Institute of Law and Technology
Veveří 70, 61180 Brno, CZ
jakub.harasta@law.muni.cz; http://cyber.law.muni.cz

Senior Assistant Professor, Masaryk University, Faculty of Law, Institute of Law and Technology
Veveří 70, 61180 Brno, CZ
matej.myska@law.muni.cz; http://cyber.law.muni.cz

***Keywords:*** ***Personal data protection, research data, secondary use, anonymization***

***Abstract:*** *Open access to research data is a growing trend, especially in the case of publicly funded research. Secondary use of personal data might be however legally challenging, especially when processing of personal data is involved. In this short paper, we deal with two issues namely use of open-ended and vague consent forms by researchers and the static perception of anonymization. To overcome these problems, we theorize a complex institutional approach employing various means to provide the needed legal certainty for secondary use of research data.*

## 1. Sharing Research Data

There are multiple motivations for sharing research data arising from various positions. As Popper noted, phenomena occurring in isolated and irreproducible fashion have no real value for scientific advancement of the human race [Popper 2002, p. 66]. This serves as the idealistic motivation for open access to research data. Beside ideals and aspirations, there is also a pragmatic approach to sharing of research data. Since it is possible to formulate and validate different research hypotheses over the same data, it is not economically efficient to collect the same data twice [Law 2005, p. 6]. Regardless of putting emphasis on idealistic or economical aspects, sharing research data supports cost-efficient funding and ensures that reproducible (and hence genuine) science takes place. In the European Union, the sharing of research data is promoted within the current Horizon 2020 Framework Programme for Research Innovation in the form of the Open Research Data Pilot.[1]

However, the public availability of research data and policies supporting its secondary use shall not interfere with protected rights of third parties, especially personal data protection. In the mentioned Open Research Data Pilot the projects beneficiaries still have the obligation to process personal data «*in compliance with applicable EU and national law on data protection*».[2] The incompatibility with rules on protecting personal data is regarded as a reason for «opt-out» of the Pilot [European Commission 2016, p. 8].

In our opinion such a bipolar «all-or-nothing» approach is not desirable. In this paper, we draw attention to some of the basic issues where personal data protection and data sharing promoting policies may clash. We also introduce and advocate a complex approach that contributes to overcoming those problematic issues and is achievable under both existing and upcoming legal frameworks.

---

[1] Art. 18(2) 32013R1291 and Art. 43(2) 32013R1291. Moreover, since 2017, the participation in this pilot is the default setting in all the thematic calls of Horizon 2020. See the European Commission Decision C(2016)4614 of 25 July 2016, EN Horizon 2020, Work Programme 2016–2017, 20. General Annexes, Part L.

[2] Art. 39 of the Horizon 2020 Model Grant Agreement, version 3.0. http://ec.europa.eu/research/participants/data/ref/h2020/mga/gga/h2020-mga-gga-multi_en.pdf (all Internet sources accessed on 10 January 2017), 2016.

## 2. Legal issues in secondary use of research data

The main challenges of the secondary use of research data are the use of open-ended and vague consent forms by researchers and the static perception of anonymization, which are not compliant with the existing legal framework.

The secondary use of research data is to be understood as use in which the data subject has only a very limited role – i.e. no consent for such use has been directly given by and obtained from the data subject. Despite this mode of use being promoted in the Open Research Data movement, the legal framework for personal data protection is generally not in favour of such use. Secondary use inherently leads to disconnection of the data subject from its data. This state of disconnection is against the *ratio* of the existing legal framework, because it makes exercising of rights of data subject much more difficult.[3] The possibility of such disconnection is directly limited by the legal requirements imposed on the consent itself – consent is required to be informed and specific.[4] Researchers often aim to fix this issue by relying on all-encompassing open consent for «*any future research activity*». However, such specification of purpose of processing personal does not meet the requirements prescribed by law.[5] The data subject is not able to assess what kind of processing is allowed under such consent. These over-inclusive consent forms, which are not eligible to allow secondary use, could be therefore seen as first problematic area.

The second conflict could be observed between the research communities' perception of anonymization and the strict legal framework dealing therewith. Anonymous data are ex-personal data [Borgesius/Gray/Van Eechoud 2015, p. 2118] that can no longer lead to identification of the data subject. Anonymization is often understood as removal of direct identifiers, which is however long known as insufficient [Dalenius 1986]. Further, the understanding of anonymization has been largely shifted by works on *k-anonymity* [Sweeney 2002] and well-known anonymization fails.[6] Anonymization has become almost impossible to achieve in the long run and even datasets claimed to be anonymous could be, as years go by, re-identified. A pragmatic approach respecting the existing state of technology was introduced to facilitate the exchange of data within a multi-national research team [Arning/ Forgó/Krügel 2009], but to our knowledge no such system exists in EU in general for secondary use of research data. The static perception of anonymization could be therefore observed as the second problematic area.

These problems could, in our opinion, be at least mitigated by employing sound organizational means to fulfil the needed legal obligations (i.e. that the rights granted to the data subjects by the effective legal regulation are respected). This system could also provide for adequate risk management related to the processing of personal data.

## 3. A complex institutional approach

Repositories of research data are becoming a standard for many research institutions,[7] as they allow centralised archiving of research data. However, these repositories are often (and especially, though not exclusively, in Central and Eastern Europe) understood in a passive way as means for external (to the public) or internal (among research teams) communication of research data. We argue and aim for repositories not only serving as static storage of research data, but for complex institutional repositories backed-up by an adequate institutional

3 Section V 31995L0046, Chapter III 32016R0679.

4 Art. 2(h) 31995L0046, Art. 4(11) 32016R0679. Moreover in the case of processing of special (sensitive) categories of data the consent must be explicit (Art. 8(2)(a) 31995L0046, Art. 9(2)(a) 32016R0679).

5 See [Hallinan/Friedewald 2015] for detailed discussion of the open consent. Article 29 Data Protecion Working Party explicitly mentions such purpose as too broad [Article 29 Data Protecion Working Party 2013, p. 52].

6 Netflix [Narayanan/Shmatikov 2008] and T3 dataset [Lewis/Kaufman/Gonzalez/Wimmer/Christakis 2010; Zimmer 2010].

7 See e.g. Harvard Dataverse (https://dataverse.harvard.edu/), University of Edinburgh DataShare (http://datashare.is.ed.ac.uk/) or the EU-repository OpenAire (https://www.openaire.eu/).

framework inciting active management of research data.[8] The general idea behind such a complex institutional approach to open research data is that the system promotes controlled secondary use by default and is not static.

Firstly, datasets that are supposed to be released for secondary use need to be managed starting from the first phase of their life cycle. Simply put, the data entering the repository should be cleared starting with its collection. Expert personnel therefore needs to be employed to guide researchers through the collection, in order not to disqualify the data from secondary use altogether. Moreover, a set of multiple consent forms must be provided and sanctioned for use for various fields of research. Consequently, the researcher considering usage of the repository must prove the use of the appropriate consent form related to the data. Without the consent the data must be discarded and not provided for secondary use.

Secondly, research data subjected to secondary use need to be either anonymized or processed with appropriate consent. Unfortunately, absolute and future-proof anonymization remains largely unachievable [Ohm 2010]. Therefore, repository and related personnel need to ensure the data intended for secondary use is processed in such a way that makes re-identification of data subjects impossible without significant costs.[9] First the checklist of direct identifiers that need to be removed must be provided.[10] The European legal framework, however, also regulates indirect identifiers.[11] Therefore, an expert assessment analysing risks of potential re-identification by aggregation of existing research data must be undertaken [Polonetsky/Omer/Kelsey 2016]. Based on such assessment, the dataset is assigned with a data tag[12] that either allows free use or prescribes for a certain level of restriction, such as a requirement of contract to be signed by the secondary user etc. Moreover, since anonymization cannot be perceived as static states anymore, but is of dynamic nature, an appropriate feedback loop is necessary to implement to allow for re-evaluation of risks associated with such potentially re-identifiable dataset. This allows for temporary or permanent take-down of the dataset or for temporary or permanent shift towards more restrictive mode of access (re-tagging of the dataset).

## 4. Conclusion and further work

The abovementioned combination of institutional guarantees (expert assessment of re-identification risks by aggregation, access modes employing data tags, feedback loop for re-evaluation of risks), contractual obligations of re-users, multiple consent forms sampled for use in different fields of research, checklists for control checklist of direct identifiers and expert oversight allows us to adjust the specificity requirement required from consent by existing legal framework and consequently control data in a legally compliant way. Our proposed solution involves setting conditions for institutional repository to carefully balance the possibility to re-use useful research data and the necessity to protect personal data.

Without such a solution, researchers will still face uncertainty on how to make the data legally available for secondary use, which in conclusion will lead to not sharing them at all [Sayogo/Pardo 2013, p. S21]. The situation is further complicated by unwillingness of many national Data Protection Authorities to incite and assist in building industry best practices or their very limited activities in this sector.[13] As the effective date of the General Data Protection Regulation approaches and Horizon 2020 sets open access to research data as the

---

8 This approach draws heavily from non-EU literature as well as EU experience and implements various conditions throughout data lifecycle. See [Ball 2002] for overview of the various data management lifecycle models.

9 For the concept of re-identification and «identifiability» of data see [Nelson 2015].

10 Similar to §164.514 (2)(i) of The Health Insurance Portability and Accountability Act of 1996 (Pub. L. 104–191, 110 Stat. 1936, available also from: https://www.law.cornell.edu/cfr/text/45/164.514) in the U.S. On the concept of personal data in the EU see the [Article 29 Data Protecion Working Party 2007].

11 Art. 2(a) 31995L0046, Art. 4(1) 32016R0679.

12 See [Sweeney/Crosas/Bar-Sinai 2015] for details on the concept of data tags and their use.

13 E.g. the Czech Data Protection Authority (Office for personal data protection) issued up until now only one Statement on the processing of personal data in the context of scienece No. 2/2006, https://www.uoou.cz/VismoOnline_ActionScripts/File.ashx?id_org=200144&id_dokumenty=9692, 2013.

default modus operandi, it is convenient time to start working with the respective Data Protection Authorities on developing a functioning best practice to secondary use of research data protection that would eliminate the bipolar all-or-nothing approach. In our opinion, sharing research data should not be an all-or-nothing choice.

## 5. References


Arning, Marian/Forgó, Nikolaus/Krügel, Tina, Data Protection in grid-based multicentric clinical trials: killjoy or confidence-building measure?, Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, 2009, volume 369, issue 1898, pp. 2729–2739. DOI 10.1098/rsta.2009.0060.

Article 29 Data Protection Working Party, Opinion 03/2013 on purpose limitation, 2 April 2014. 00569/13/EN WP 203.

Article 29 Data Protection Working Party, Opinion 4/2007 on the concept of personal data, 20 June 2007. 01248/07/EN WP 136.

Ball, Alexander, Review of data management lifecycle models, 2012. http://opus.bath.ac.uk/28587/.

Borgesius, Frederik Zuiderveen/Gray, Jonathan/Van Eechoud, Mireille, Open Data, Privacy, and Fair Information Principles: Towards a Balancing Framework, Berkeley Technology Law Journal, 2015, volume 30, issue 3, pp. 2073–2131. DOI 10.15779/Z389S18.

Dalenius, Tore, Finding a Needle in a Haystack or Identifying Anonymous Census Records, Journal of Official Statistics, 1986, volume 2, issue 3, pp. 329–336.

European Commission, H2020 Programme Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020, 25 August 2016, Version 3.1. http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf.

Hallinan, Dara/Friedewald, Michael, Open consent, biobanking and data protection law: can open consent be «informed» under the forthcoming data protection regulation?, Life Sciences, Society and Policy, 2015, volume. 11, issue 1. DOI 10.1186/s40504-014-0020-9.

Law, Margaret, Reduce, Reuse, Recycle: Issues in the Secondary Use of Research Data, IASSIST quarterly, 2005, volume 29, issue 1, pp. 5–10. http://www.iassistdata.org/sites/default/files/iqvol291law.pdf.

Lewis, Kevin/Kaufman, Jason/Gonzalez, Marco/Wimmer, Andreas/Christakis, Nicholas, Tastes, ties, and time: A new social network dataset using Facebook.com, Social Networks, 2008, volume 30, issue 4, pp. 330–342. DOI 10.1016/j.socnet.2008.07.002.

Narayanan, Arvind/Shmatikov, Vitaly, Robust De-anonymization of Large Sparse Datasets. In: Proceedings of the 2008 IEEE Symposium on Security and Privacy, IEEE, 2008, pp. 111–125.

Nelson, Gregory S., Practical Implications of Sharing Data: A Primer on Data Privacy, Anonymization, and De-Identification, Paper 1884-2015, ThotWave Technologies, Chapel Hill, NC. https://www.thotwave.com/wp-content/uploads/2015/09/data_sharing_privacy_anonymization_and_de-identification_rev_13.pdf.

Ohm, Paul, Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization, University of California Law Review, 2010, volume 57, issue 6, pp. 1701–1778.

Polonetsky, Jules/Omer, Tene/Kelsey, Finch, Shades of Gray: Seeing the Full Spectrum of Practical Data De-Identification, Santa Clara Law Review, 2016, volume 56, issue 3, pp. 593–629.

Popper, Karl, The Logic of Scientific Discovery, Routledge, London 2002.

Sayogo, Djoko Sigit/Pardo, Theresa A., Exploring the determinants of scientific data sharing: Understanding the motivation to publish research data, Government Information Quarterly, 2013, volume 30, supplement 1, pp. S19–S31. DOI 10.1016/j.giq.2012.06.011.

Sweeney, Latanya, k-Anonymity: A Model for Protecting Privacy, International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 2002, volume 10, issue 5, pp. 557–570.

Sweeney, Latanya/Crosas, Mercé/Bar-Sinai, Michael, Sharing Sensitive Data with Confidence: The Datatags System, Technology Science, 2015. http://techscience.org/a/2015101601.

Zimmer, Michael, «But the data is already public»: on the ethics of research in Facebook, Ethics and Information Technology, 2010, volume 12, issue 4, pp. 313–325. DOI 10.1007/s10676-010-9227-5.